A Bayesian Zero-One Inflated Beta Model for Estimating Poverty in U.S. Counties *

Jerzy Wieczorek Sam Hawala†

Abstract

We propose and evaluate a Bayesian beta regression model for U.S. county poverty rates. Such a rate model could be an improvement to the U.S. Census Bureau's current small-area poverty approach of linearly modeling the logarithm of poverty levels. For small areas, some of which may have estimates of no poverty or all poverty, a zero-one inflated rate model can usefully account for estimated rates of 0 or 1. Using Bayesian computation techniques, we estimate the parameters of a zero-one inflated beta regression model. We compare the results to the Census Bureau's current small-area model for county poverty estimation.

Key Words: small area estimates, SAIPE, MCMC, beta regression, hierarchical model

1. Introduction

The Small Area Income and Poverty Estimates (SAIPE) program at the U.S. Census Bureau uses small area estimation techniques to create model-based estimates of selected poverty and income statistics. The estimates are intended to be more timely than direct estimates from the decennial census or five-year American Community Survey (ACS), as well as more precise and stable than single-year ACS direct estimates for small areas.

In this paper, we are concerned with estimating the number of related poor children aged 5-17 in U.S. counties. These estimates are provided to the Department of Education and used for allocating federal funding to local programs. In 1998 a panel of the National Research Council studied alternative county models. In its report, the panel deems the county model to be at "the heart of the estimation procedure that develops estimates of school-age children in poverty to allocate federal funds under Title I of the Elementary and Secondary Education Act for education programs to aid disadvantaged children."

The existing county-level approach is based on a Fay-Herriot 'log-level' model, i.e. a model on the natural log of the number of related poor children in each area. The model combines single-year ACS direct estimates with regression predictors from administrative data records including Internal Revenue Service (IRS) tax data and Supplemental Nutrition Assistance Program (SNAP) ('food stamp') data. This Fay-Herriot model is described in more detail on the Small Area Income and Poverty Estimates website (U.S. Census Bureau, 2010). The current SAIPE model is tractable and well-established, but it is worth considering alternative models that may have advantages over the current approach.

In particular, some counties have ACS direct estimates of zero related children in poverty. Since log(0) is undefined, these counties must be dropped from the estimation procedure, with a resulting loss of information and efficiency. During

*This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress. Any views expressed on statistical, methodological, technical, or operational issues are those of the authors and not necessarily those of the U.S. Census Bureau.

†Social, Economic, and Housing Statistics Division, U.S. Census Bureau, 4600 Silver Hill Road, Washington, DC 20233

four out of the five years from 2005 to 2009, over 5% of counties had ACS direct estimates of zero related children in poverty.

Furthermore, Census Bureau staff have found other concerns with the log-level model, including biased direct variance estimates on the log scale (Huang and Bell, 2009), and have suggested considering modeling poverty rates rather than poverty counts.

The model we propose to account for both of these issues is a zero-one inflated Beta (ZOIB) regression model. The Beta distribution allows us to model poverty rates directly on a continuous range between 0 and 1 exclusive; and a multinomial component allows us to add the 0 and 1 endpoints to our model. The ZOIB model details are discussed in section 2.

This hierarchical model is difficult to solve by classical/analytical methods but lends itself well to Bayesian treatment by Markov Chain Monte Carol (MCMC) methods. Furthermore, posterior distributions give us a lot of information and allow for useful model-checking approaches. The computational details are discussed in section 3.

Finally, section 4 presents our model evaluation results, which we briefly summarize here:

- We conduct a simulation study to check coverage rates for credible intervals (CIs) (the Bayesian analogue of confidence intervals) in order to evaluate whether the MCMC approach is working, finding that our 95% CIs do indeed cover the true value around 95% of the time.
- We summarize posterior point estimates and compare them to ACS direct estimates as well as SAIPE estimates under the current official model.
- We compare the predicted expected frequency of 0s and 1s to the number actually observed in the data, finding that they are similar.
- We partition counties into several groups, by county size, and evaluate the average mean square error (AMSE) for each group of counties, finding that the AMSEs are of similar order of magnitude to the current SAIPE model AMSEs for the largest counties, while smaller counties still have some room for improvement under our model.

2. Model

We begin the model description in section 2.1 by discussing the simplified case where no 0s or 1s are observed, in which case a beta regression can be a suitable model for the observed poverty rates. Next, in section 2.2, we describe the full zero-one inflated beta (ZOIB) model which allows for observed rates of 0 or 1.

2.1 Beta regression model

First let us consider the case where all observed estimated county-level poverty rates (among related children aged 5-17) are in the open interval $(0, 1)$.

Let Y_i denote the true poverty rate in county i and let y_i denote the observed estimated poverty rate in county i. We define y_i as

$$y_i = \frac{\text{ACS estimated number in poverty}}{\text{ACS estimated number in poverty universe}}$$

and assume that all rates fall within the interval $(0, 1)$.

A reasonable candidate for modeling y is the beta distribution, which covers a wide range of distribution shapes on the interval $(0, 1)$. The density function of the beta distribution (with parameters a and b) is

$$p(y|a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} y^{a-1}(1-y)^{b-1}$$

The estimated poverty rate is also related to four covariates (tax poverty rate, tax non-filing rate, food stamp participation rate, and natural log of ACS sample size in poverty universe) that we observe for each county, using administrative and ACS records (U.S. Census Bureau, 2011). In order to incorporate this covariate information into a beta regression model, we parameterize the beta family in terms of its mean,

$$\mu = E(y) = \frac{a}{a+b},$$

and a parameter related to its variance,

$$\gamma = a + b$$

Inversely, the parameters a, b can be expressed as $a = \gamma\mu$ and $b = \gamma(1-\mu)$. Note that the variance of a beta distribution is

$$Var(y) = \frac{\mu(1-\mu)}{\gamma+1} = \frac{ab}{(a+b)^2(a+b+1)}$$

The variance does depend on the mean μ, and larger values of γ correspond to less heterogeneity in the data.

Now we can model poverty rates as a beta regression, with each county having its own μ_i and γ_i.

Furthermore, we can rewrite $Var(y)$ as a variance of rate estimates rescaled by a design effect:

$$Var(y_i) = \frac{\mu_i(1-\mu_i)}{\gamma_i+1} = \frac{\mu_i(1-\mu_i)}{n_i}\text{deff}_i$$

where n_i is the (unweighted) size of the sample in the poverty universe in county i. This leads to

$$\gamma_i = \frac{n_i}{\text{deff}_i} - 1$$

The design effect reflects the effect of the complex sample design (Kish, 1965). For now, we approximate deff_i by an approximate design effect for a stratified simple random sample with a negligible sampling fraction in all strata.

$$\text{deff}_i = n_i \sum_h \hat{W}_{ih}^2 / n_{ih} \tag{1}$$

$$\hat{W}_{ih} = \hat{N}_{ih}/\hat{N}_i, \qquad \hat{N}_{ih} = \sum_{c\in h} w_{ic}, \qquad \hat{N}_i = \sum_{h\in i} \hat{N}_{ih}, \qquad n_i = \sum_h n_{ih}$$

w_{ic} is the weight for child c in household h, which is in area i. The $\hat{W}_{ih}$, $\hat{N}_{ih}$, and $\hat{N}_i$ are sample estimates of the corresponding population quantities.

Liu et al. (2001) discuss this approach using population quantities

$$N_{ih}, \quad N_i = \sum_{h\in i} N_{ih}, \quad \text{and} \quad W_{ih} = N_{ih}/N_i \tag{2}$$

for small area estimation. Kalton et al. (2005) discuss it for estimating the design effect for a disproportionate stratified sample. They derive the estimate for the design effect under the assumptions of equal stratum means and equal within-stratum variances. They illustrate that, for proportions between 0.2 and 0.8, the formula using (2) to estimate deff_i provides reasonably close estimates of the design effect.

For now, the deff_i are used in formula (1) to estimate the γ_i, which are then treated as known for the remainder of the estimation procedure. Future work may involve using the modeled poverty estimates to re-estimate the γ_i and vice versa, iteratively, until both the poverty rates and the γ_i stabilize.

As given above, γ_i is undefined for areas where no children in the poverty universe were sampled. Our current procedure drops these counties from the estimation procedure.

Let $\boldsymbol{\beta_\mu}$ denote a vector of regression coefficients (including an intercept) for estimating μ_i, so that $\boldsymbol{\beta_\mu} = (\beta_{\mu 0}, \beta_{\mu 1}, \ldots, \beta_{\mu 4})$. Also let $\mathbf{y} = (y_1, \ldots, y_m)$ be the vector of observed ACS direct estimates for each of the m small areas, and let the corresponding covariate vectors be $\mathbf{x}_i = (1, x_{i1}, \ldots, x_{i4})$ for each $i = 1, \ldots, m$.

The Bayesian beta "regression" model we consider is

$$\begin{aligned} y_i | \mu_i, \gamma_i, \mathbf{x}_i, \boldsymbol{\beta_\mu} &\sim \text{Beta}\left(\gamma_i \mu_i, \gamma_i (1 - \mu_i)\right) \\ \text{logit}\left(\mu_i\right) = \text{logit}\left(\mu(\mathbf{x}_i')\right) &= \mathbf{x}_i' \boldsymbol{\beta_\mu} \\ p(\boldsymbol{\beta_\mu}) &\propto 1 \end{aligned}$$

where $p(\boldsymbol{\beta_\mu})$ is an improper flat prior on the $\boldsymbol{\beta_\mu}$ vector. We use the logit link $F(w) = \log(\frac{w}{1-w})$, with inverse link function $F^{-1}(w) = \frac{\exp(w)}{1+\exp(w)}$.

Our model's parametric structure specifies a linear relationship (on the logit scale) between μ, x, and β_μ. There are several ways to broaden the scope of the models we can consider. One is to use a semiparametric structure, where $\text{logit}(\mu)$ will be related to x via an arbitrary function. Another is to impose a stochastic model such as $\text{logit}(\mu_i) \sim \text{Normal}(\mathbf{x}_i' \boldsymbol{\beta_\mu}, \sigma^2)$. However, we have had difficulties getting the MCMC code to converge for that model.

2.2 Zero-one inflated beta model

In practice, we do observe ACS direct estimates of the poverty rate of exactly 0 or exactly 1. In other words, our model should ultimately allow for observations in the closed interval $[0, 1]$.

Let Z_i denote the true poverty rate in county i and let z_i denote the observed estimated poverty rate in county i. Rates may be in the interval $[0, 1]$. Let $p_i^{(0)}$ be the probability that county i has an observed rate of 0, and similarly let $p_i^{(1)}$ be the probability of observing 1. Otherwise, the county has a probability of $p_i^{(01)} = (1 - p_i^{(0)} - p_i^{(1)})$ of having a rate drawn from the $\text{Beta}(a_i, b_i)$ distribution.

In other words, the result of a multinomial trial is used to determine which of the three processes generates an observation.

$$z_i = \begin{cases} 0 & \text{with probability } p_i^{(0)}, \\ 1 & \text{with probability } p_i^{(1)}, \\ \sim \text{Beta}(a_i, b_i) & \text{with probability } p_i^{(01)}. \end{cases}$$

$\boldsymbol{\beta_\mu}$ is the vector of regression coefficients for estimating the logit of the mean of the beta distribution, $\mu_i = a_i/(a_i + b_i)$. Similarly, we model $p_i^{(0)}$ and $p_i^{(1)}$ using $\boldsymbol{\beta_0}$

and $\boldsymbol{\beta_1}$:

$$p_i^{(0)} = \text{inv.logit}(\boldsymbol{x_i}'\boldsymbol{\beta_0}) \quad \text{and} \quad p_i^{(1)} = \text{inv.logit}(\boldsymbol{x_i}'\boldsymbol{\beta_1})$$

We use an intercept term and the same set of four regressors as before (tax poverty rate, tax non-filing rate, food stamp participation rate, and natural log of ACS sample size in poverty universe) for each of $\boldsymbol{\beta_0}$, $\boldsymbol{\beta_1}$, and $\boldsymbol{\beta_\mu}$.

The random variable z_i in this zero/one-inflated model has the following mean:

$$E(z_i) = 0 \cdot p_i^{(0)} + 1 \cdot p_i^{(1)} + \mu_i \cdot (1 - p_i^{(0)} - p_i^{(1)}) = p_i^{(01)}\mu_i + p_i^{(1)}$$

Note that we make the assumption that the ACS direct estimate z_i is unbiased for the true poverty rate that we are trying to estimate. Therefore, despite the structure of the model, our estimate $\hat{Z}_i$ of the poverty rate is not the estimated mean of the beta distribution, $\hat{\mu}_i$, but rather the estimated posterior mean of z_i.

In our MCMC procedure we draw samples $j = 1, ..., J$ from the posteriors of each parameter. Letting tildes denote these sampled posterior draws, we have

$$\begin{aligned} \tilde{Z}_{ij} &= \tilde{p}_{ij}^{(01)}\tilde{\mu}_{ij} + \tilde{p}_{ij}^{(1)} \\ \hat{Z}_i &= J^{-1}\sum_{j=1}^{J}\tilde{Z}_{ij} \end{aligned} \tag{3}$$

3. Computational Approach

We use the Metropolis algorithm to generate a MCMC sample from the posterior distribution.

First, propose new values for the β values (regression coefficients) for each of the three parts of the model. (To do this, use a jumping distribution to move from the previous accepted β values to a new proposal value, and just tune this procedure over many iterations to give an appropriate acceptance rate. We use independent normal jumping distributions for each β in the first burn-in period, then use the sample covariance matrix of previous iterations to choose a multivariate normal jumping distribution for the second burn-in period.)

For each iteration, use the proposed β values to calculate μ_i and $p_i^{(0)}$ and $p_i^{(1)}$ for each county and convert into a_i and b_i for each county. Then plug the values into the joint posterior likelihood function, which is proportional to

$$\begin{aligned} &p(\boldsymbol{\beta_0}, \boldsymbol{\beta_1}, \boldsymbol{\beta_\mu} | \boldsymbol{\gamma}, \mathbf{z}, \mathbf{x}) \propto \\ &p(\boldsymbol{\beta_0}, \boldsymbol{\beta_1}, \boldsymbol{\beta_\mu}) \prod_{i=1}^{m} p_{0i}^{I\{z_i=0\}} p_{1i}^{I\{z_i=1\}} [p_{01i}\text{pdf}_{\text{Beta}}(z_i|a_i, b_i)]^{I\{0<z_i<1\}} \end{aligned}$$

where $p(\boldsymbol{\beta_0}, \boldsymbol{\beta_1}, \boldsymbol{\beta_\mu}) \propto 1$ is an improper flat prior on the $\boldsymbol{\beta}$ vectors, and

$$p_{0i} = \text{inv.logit}(\boldsymbol{x_i}'\boldsymbol{\beta_0}), \quad p_{1i} = \text{inv.logit}(\boldsymbol{x_i}'\boldsymbol{\beta_1}), \quad \mu_i = \text{inv.logit}(\boldsymbol{x_i}'\boldsymbol{\beta_\mu})$$
$$a_i = \gamma_i\mu_i, \quad b_i = \gamma_i(1 - \mu_i)$$

Compare to the likelihood from the previous iteration. If either the ratio of new-to-old likelihoods is higher than 1, or if it's smaller but a random uniform $(0, 1)$ draw is even smaller, then accept the new values. Otherwise, reject the new values; store the old values as the values for this iteration; and draw a new proposal values for the next iteration. The decision to accept or reject a set of proposed values is made for all three $\boldsymbol{\beta}$ vectors simultaneously.

Gelman et al. (2004) suggest adjusting the proposal distributions to keep the acceptance rate around 20-40%. Once the rate is steadily in that range, draw the desired number of samples, discard the burn-in period, and make inferences based on the rest.

We do all of this for several chains at once (starting with different, overdispersed initial values). This way, we can evaluate the convergence by checking for each β whether its multiple chains are converging to the same distribution, using the potential scale reduction factor $\hat{R}$ (Gelman et al., 2004, p. 297).

Finally, we compute our estimates of the rate in each county by following the equations in (3). For county i, transform x_i and all of the post-burn-in samples of $\tilde{\boldsymbol{\beta}}_{\mathbf{0}\boldsymbol{j}}$, $\tilde{\boldsymbol{\beta}}_{\mathbf{1}\boldsymbol{j}}$, and $\tilde{\boldsymbol{\beta}}_{\boldsymbol{\mu}\boldsymbol{j}}$ into samples of $\tilde{p}_{ij}^{(0)}$, $\tilde{p}_{ij}^{(1)}$, and $\tilde{\mu}_{ij}$ and consequently $\tilde{Z}_{ij}$. Then average the $\tilde{Z}_{ij}$ samples to estimate the posterior mean $\hat{Z}_i$.

4. Results

This section summarizes the poverty estimates from our procedure and several evaluations of those estimates, as well as a simulation study designed to evaluate the estimation procedure itself.

The first two sections refer to the simulation study. Section 4.1 explains how the simulation study datasets were generated, and section 4.2 describes the estimation procedure's credible interval coverage under the simulation study assuming the model is true.

The remaining sections summarize results based on the actual data. Section 4.3 summarizes the parameter estimates using standardized posterior means and 95% credible intervals. Section 4.4 compares the observed number of ones and zeros to the model's predicted number of ones and zeros. Section 4.5 summarizes the residual differences between our estimates and the ACS direct estimates, overall as well as after sorting the counties into groups by population size. Finally, section 4.6 shows the estimated average MSEs (AMSEs) within each of the county groups.

4.1 Details of simulation study

Before simulating new datasets, we began by using the 2009 SAIPE input data to find the MCMC posterior estimates of the $\boldsymbol{\beta}$ vectors under the model in section 2.2. (These estimates are displayed in section 4.3.) Keeping the same X matrix and $\boldsymbol{\gamma}$ vector from the 2009 data, and treating the posterior estimates of the $\boldsymbol{\beta}$ vectors as a fixed set of known parameter values, we simulated 50 new datasets of $\boldsymbol{z}$ based on the model.

Our approach assumed that the sample size and design effect, and therefore also γ_i, remain constant in each county across the simulations. Hence the survey weights were only used for the initial estimation of the design effect based on the actual observed ACS estimates.

In other words, the fixed X matrix and set of $\boldsymbol{\beta}$ vectors defined the 'true' values of $\boldsymbol{p}^{(0)}$, $\boldsymbol{p}^{(1)}$, and $\boldsymbol{\mu}$. These were used to define the 'true' values of the poverty rates $\boldsymbol{Z}$, as well as to simulate new draws from the Beta distribution for each county and then the multinomial choice of whether to use 0, 1, or the Beta draw. We used this approach to generate 50 new simulated datasets for each county, keeping the $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ vectors fixed across the 50 simulations.

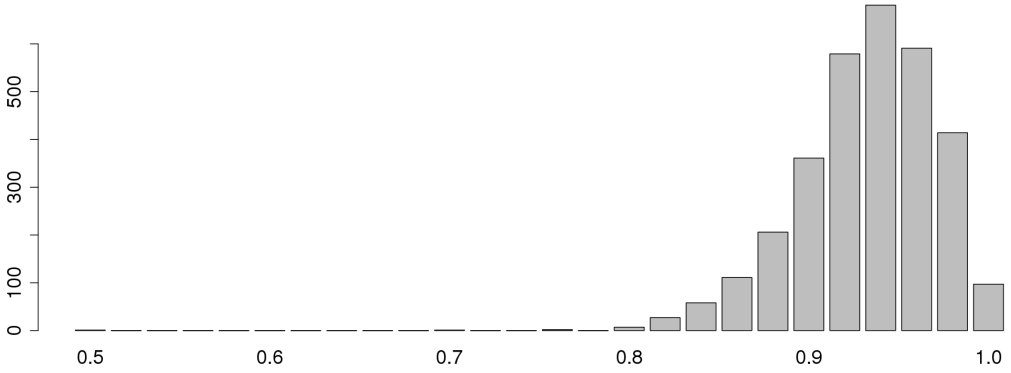

Figure 1: Number of counties with each CI coverage rate, for 95% CIs computed over 50 simulations
Source: Simulated from U.S. Census Bureau, Small Area Income and Poverty Estimates data, 2009

4.2 Credible interval coverage

We used the simulated datasets described in section 4.1 to evaluate the MCMC's credible interval (CI) coverage, assuming the model in section 2.2 is true.

We ran the estimation procedure on each of the 50 simulated datasets to see how often each county's 95% CI for Z_i contains the true value. For a given dataset, each county's 95% CI is determined by taking the set of J posterior samples $\tilde{Z}_{ij}$ and using the 2.5% and 97.5% percentiles of the samples as the upper and lower credible limits, respectively. For each dataset and each county, we note whether the true value Z_i falls within these limits.

In our 50 simulated datasets, most counties' CIs contained the true value close to 95% of the time. To be precise, the most frequent CI coverage was 94%, i.e. 47 of the 50 datasets' CIs contained the true value. This happened in 681 of the 3136 counties in the dataset. See Figure 1 for the full histogram.

4.3 Parameter estimates

The standardized posterior means and 95% credible intervals for the $\boldsymbol{\beta}$ parameters are summarized in Table 1. The posterior means and credible interval endpoints for each parameter were standardized by dividing them by the parameter's posterior standard deviation.

Given that we only observed a single estimate of 1, the $\hat{\boldsymbol{\beta}}_{\mathbf{1}}$ estimates are unstable on this 2009 dataset, even once the MCMC chains appeared to reach convergence. However, the $\hat{\boldsymbol{\beta}}_{\mathbf{0}}$ estimates are much less variable, with only the non-filing rate predictor's CI containing zero. The $\hat{\boldsymbol{\beta}}_{\boldsymbol{\mu}}$ estimates have the tightest CIs and appear quite stable.

The coefficient estimates support two expected conclusions: First, the negative posterior means of the coefficients of $\ln(n)$ for $\hat{\boldsymbol{\beta}}_{\mathbf{0}}$ and $\hat{\boldsymbol{\beta}}_{\mathbf{1}}$ suggest that the probability of observing a 0 or 1 is lower for larger counties. Second, the posterior means of the tax poverty rate, non-filing rate, and food stamp participation rate coefficients for $\hat{\boldsymbol{\beta}}_{\boldsymbol{\mu}}$ are all positive, suggesting that poverty is higher in counties with high values of these predictors.

	Intercept	TaxPov	NonFiler	FoodStamp	ln(n)
Beta0 Mean	9.47	-2.78	0.07	-4.63	-15.29
Beta0 2.5%	7.54	-4.94	-1.92	-6.58	-17.39
Beta0 97.5%	11.54	-0.85	1.96	-2.71	-13.36
Beta1 Mean	-1.80	0.74	1.64	-1.17	-1.08
Beta1 2.5%	-3.74	-1.40	-0.15	-3.47	-3.37
Beta1 97.5%	0.02	2.67	3.74	0.42	0.52
BetaMu Mean	-147.27	50.91	14.12	24.71	20.16
BetaMu 2.5%	-149.25	49.00	12.07	22.68	18.20
BetaMu 97.5%	-145.32	52.93	16.03	26.70	22.12

Table 1: Standardized posterior means and 95% credible intervals for $\boldsymbol{\beta}$
Source: Simulated from U.S. Census Bureau, Small Area Income and Poverty Estimates data, 2009

4.4 Checking predicted vs. actual estimates of ones and zeros

We would like to see whether the model predicts as many 0s and 1s as were actually observed in the data. Using the same data as the SAIPE 2009 estimation procedure (2009 ACS data with 2008 tax and SNAP administrative data), the model predicts $\sum \hat{p}_i^{(0)} = 173.9$ zeros and $\sum \hat{p}_i^{(1)} = 0.066$ ones overall. This is comparable to the 174 zeros observed in the ACS estimates, but perhaps somewhat low for the single ACS observation of value one.

We bin the number of zeros by the ACS poverty universe sample size in a county to see if the model accurately predicts the numbers of zeros across the range of sample sizes. The sample size bins are 1-5, 6-10, 11-20, 21-50, and over 50. Results are presented in Table 2. (We do not do this for the 1s since there was only a single observed estimate of 1.)

County sample size	<5	6-10	11-20	21-50	>50
Number of zeros observed	40.0	32.0	47.0	42.0	13.0
Number of zeros predicted	38.0	36.0	47.2	39.7	12.9

Table 2: Observed and predicted number of zeros, by sample size in county
Source: Simulated from U.S. Census Bureau, Small Area Income and Poverty Estimates data, 2009

Within each of these sample size categories, the number of predicted 0s is close to the number of 0s observed. This suggests that the ZOIB model does a good job of modeling the 0s.

4.5 Residuals for posterior point estimates

Table 3 compares our point estimates of $\hat{Z}_i$ to observed z_i values from the ACS. The interquartile range shows that many differences are small, only about $\pm$ 5 percentage points, but the minimum and maximum differences are rather large, showing that a few counties' model estimates may be quite far from the ACS direct estimate.

Figure 2 shows boxplots of the residuals $\hat{Z}_i - z_i$ within each county size group, with a separate boxplot on the far left for counties with observed values of 0 or 1. (Of the 174 0s, 144 occurred in counties with populations under 10k; 20 occurred in counties of size 10-20k; and 10 occurred in counties of size 20-65k; the only 1 occurred in a county of population below 10k). Positive residuals are overestimates

	$\hat{Z}_i - z_i$
Min.	-0.894
1st Qu.	-0.055
Median	0.000
Mean	-0.012
3rd Qu.	0.048
Max.	0.646

Table 3: Summary of residuals: model estimate $\hat{Z}_i$ minus direct estimate z_i
Source: Simulated from U.S. Census Bureau, Small Area Income and Poverty Estimates data, 2009

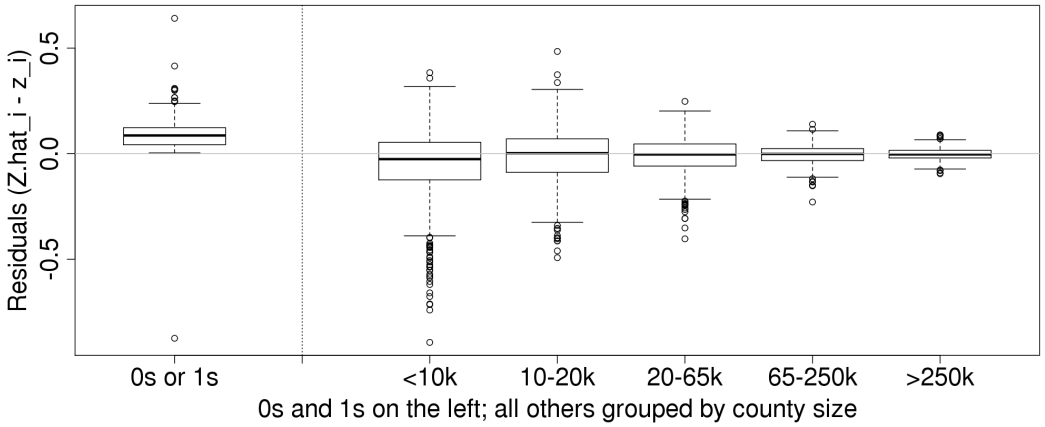

Figure 2: Residuals $\hat{Z}_i - z_i$ within each group of counties,
partitioned by county population size (counties with obs. of 0 or 1 separated out)
Source: Simulated from U.S. Census Bureau, Small Area Income and Poverty Estimates data, 2009

(i.e. the ZOIB estimates are higher than the direct estimates), and vice versa for negative residuals.

It is clear that the smallest counties' residuals are most variable, while ZOIB estimates for the largest counties tend to stay close to the ACS estimates, as expected. The far-left boxplot shows that many of the estimates for counties with observed 0s are still not too far above 0. However, for the single county with an observation of 1 (the negative residual in the far-left boxplot), the ZOIB estimate is quite far from 1. In a year with more observations of 1, we would be likely to have more stable parameter estimates and poverty rate estimates for such counties.

4.6 AMSEs for poverty rate, by county size

For internal evaluations of the current SAIPE log-level model, counties are partitioned into groups by county population size and the average MSE (AMSE) is estimated within each group. For the 2009 data, the five groups (and the number of counties in each group) are: population under 10k (n=705), 10k to 20k (n=615), 20k to 65k (n=1024), 65k to 250k (n=537), and over 250k (n=255).

We calculate similar AMSE estimates for the ZOIB model. However, unlike the official SAIPE model, the current version of the ZOIB model does not include a model error term and involves no raking. Thus the AMSEs may not be measuring exactly the same thing here as in the current SAIPE model. Nonetheless, we believe

that comparing the AMSE estimates for ZOIB vs. log-level SAIPE can give a rough sense of how much improvement remains to be made to the ZOIB model.

Our AMSE estimator for each of the 5 groups is based on the Gonzalez-Waksberg estimator, as presented e.g. in Lahiri and Pramanik (2010). For group g,

$$\text{AMSE}_g^{GW} = n_g^{-1} \sum_{i \in g} (\hat{Z}_i - z_i)^2 - n_g^{-1} \sum_{i \in g} V(z_i)$$

where $V(z_i)$ is the ACS estimate of the sampling variance of the ACS direct estimate z_i.

However, while the usual Gonzalez-Waksberg estimator averages the $V(z_i)$ values, the SAIPE team's internal evaluations take their median instead. So in fact we use an alternate estimator:

$$\text{AMSE}_g^{GW.alt} = n_g^{-1} \sum_{i \in g} (\hat{Z}_i - z_i)^2 - \text{median}_{i \in g}(V(z_i))$$

This is because counties with direct estimates of 0 or 1 have no sampling error estimates, so the mean would be undefined. The median is seen as more robust than the mean when dropping undefined counties, and the median is what tends to be used for SAIPE program internal evaluations, so we use it here for compatibility.

Finally, within each group of counties, the AMSE is usually transformed into a root AMSE on the percentage point scale. Root AMSE estimates for the ZOIB model results are shown in Table 4, along with the same estimates for the current SAIPE model.

County population size	<10k	10-20k	20-65k	65-250k	>250k
ZOIB root AMSE	13.9	9.1	5.9	3.3	2.9
SAIPE root AMSE	7.6	7.5	2.9	—	—

Table 4: Estimates of Root AMSE (on percentage point scale)
of ZOIB and SAIPE model estimates within each group of counties
Source: Simulated from U.S. Census Bureau, Small Area Income and Poverty Estimates data, 2009

For the current SAIPE model on the 2009 data, in each of the two groups of largest counties, the median sampling variance is bigger than the mean squared residual. Hence, the AMSEs are negative and the root AMSEs are undefined.

For the ZOIB model estimates in the largest counties, the root AMSE of 2.9 is the same as that found for the current SAIPE model estimates in medium-sized counties. The root AMSEs for smaller counties under ZOIB are consistently larger than for the current SAIPE model, showing that our ZOIB model still needs improvement.

However, the AMSE seems very sensitive to the choice of mean vs. median, and the possibility of negative AMSE estimates is undesirable. We intend to find a more robust measure of comparison to use for future work.

5. Future Research

In the future, the most important improvement is to add a model error term. In other words, the logits of $p^{(0)}$, $p^{(1)}$, and μ should have a stochastic distribution around the regression means $\mathbf{x}_i'\boldsymbol{\beta}$ rather than being deterministically related. Currently we are having trouble with MCMC convergence in this regard, but we hope to overcome the problem soon. This will transform our approach from a sample

regression model to a Fay-Herriot style area-level model that shrinks survey estimates towards regression predictions according to the reliability of each component. Further work should also assess the effect of benchmarking to state estimates and the value of using alternative regressors.

Another clear area for improvement is the estimation of the design effect. An updated approach is being developed. We will also consider including predictors for sampling error variances.

We will also investigate the use of replacing the improper flat priors with more informative priors. For example, prior years of ACS data and SAIPE estimates should provide ample information that we can incorporate into informative priors on the regression coefficients $\boldsymbol{\beta}$ as well as on any model error terms that may be added to the model.

REFERENCES

Huang, E.T. and Bell, W.R. (2009), "A Simulation Study of the Distribution of Fays Successive Difference Replication Variance Estimator," *Proceedings of the American Statistical Association, Survey Research Methods Section, [CD-ROM]*, Alexandria, VA: American Statistical Association.

Gelman, A., Carlin, J., Stern, H., and Rubin, D. (2004), *Bayesian Data Analysis*, Boca Raton: Chapman & Hall.

Hawala, S., and Lahiri, P. (2010), "Variance Modeling in the U.S. Small Area Income and Poverty Estimates Program for the American Community Survey," *Proceedings, Section on Bayesian Statistical Science, Section on Survey Research Methods, AS[CD-ROM]*, Alexandria, VA: American Statistical Association.

Kalton, G., Brick. J.M., and Le, T. (2005), "Estimating components of design effects for use in sample design," *Household Sample Surveys in Developing and Transition Countries*, Chapter VI, New York, NY: United Nations Statistics Division.

Kish, L. (1965), *Survey Sampling.* New York: John Wiley.

Lahiri, P., and Pramanik, S. (2010), "Estimation of Average Design-based Mean Squared Error of Synthetic Small Area Estimators," *Statistics Canada International Symposium Series: Proceedings*, Ottawa, Ontario: Statistics Canada.

Liu, B., Lahiri, P., and Kalton, G. (2007), "Hierarchical Bayes Modeling for Survey-Weighted Small Area Proportions," *Proceedings of the Section on Survey Research Methods*, Alexandria, VA: American Statistical Association.

National Research Council (1998), "Small-Area Estimates of School-Age Children in Poverty," Constance F. Citro, Michael L. Cohen, and Graham Kalton, eds. Panel on Estimates of Poverty for Small Geographic Areas, Committee on National Statistics. Washington, D.C.: National Academy Press.

U.S. Census Bureau (2010), "2006 - 2009 County-Level Estimation Details," last modified December 8, 2010, http://www.census.gov/did/www/saipe/methods/statecounty/20062009county.html

U.S. Census Bureau (2011), "Information about Data Inputs," last modified March 14, 2011, http://www.census.gov/did/www/saipe/data/model/info/index.html